Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Europ. J. Agronomy 23 (2005) 170–182

European Journal of Agronomy

www.elsevier.com/locate/eja

Integrated viewing and analysis of phenotypic, genotypic and environmental data with “GenPhEn arrays”

Jeffrey W. White[a,*], Gerrit Hoogenboom[b]

[a] USDA-ARS, U.S. Water Conservation Laboratory, 4331 E. Broadway Rd., Phoenix, AZ 85040-8834, USA
[b] Department of Biological and Agricultural Engineering, College of Agricultural and Environmental Sciences, University of Georgia, Griffin, GA 30223-1797, USA

Received 30 June 2004; received in revised form 22 October 2004; accepted 15 November 2004

Abstract

A recurring challenge in agronomic research is how to interpret large data sets that combine information on genotypes, phenotypes and environments. High-resolution color graphics offer the possibility of presenting such data as pseudo-maps or arrays, where *x*- and *y*-coordinates represent genotypes and environments, and the *z*-values represent phenotypic data using dots or other symbols and an appropriate color scheme to indicate the range of values. This paper describes use of such “GenPhEn arrays” with data from three studies: a survey of leaf thickness in common bean (*Phaseolus vulgaris* L.) lines, a multi-location trial of wheat (*Triticum aestivum* L.) lines and a simulation analysis for response of common bean to increased air temperature. By standardizing the phenotypic values, the genotypic and environmental effects can be easily viewed and better comprehended, especially when presented in multi-trait arrays. The arrays allow presenting large amounts of data (i.e., 5000 data points or more) in a compact and readily interpretable fashion. Appending symbols to indicate levels of significance of specific effects or to characterize genotypes or environments can further assist interpretation and hypothesis generation. It is expected that GenPhEn arrays can be used by plant breeders, agronomists and others to rapidly examine large sets of data for patterns that merit further study using more quantitative approaches.
Published by Elsevier B.V.

Keywords: Bioinformatics; Data visualization; Environment; Modeling; Plant breeding; Phenotype; Statistical analysis

1. Introduction

Research that is concerned with how crop phenotypes respond to genotypic and environmental manipulation typically requires examining effects of multiple crop traits and environmental factors and their interactions. Data sets combining information on genotypes, phenotypes and environments are often large and complex, and the quantity and complexity of data can be expected to increase with advances in field instrumentation, molecular biology and other scientific tools. Statistical approaches such as multiple regressions (e.g.,

* Corresponding author. Tel.: +1 602 437 1702x268; fax: +1 602 437 5291.
E-mail address: jwhite@uswcl.ars.ag.gov (J.W. White).

1161-0301/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.eja.2004.11.005

White et al., 1992), principal component analysis (Yan et al., 2001), stability analysis (Lin et al., 1986) and cluster analysis (Collaku et al., 2002) are used widely to generate numerical parameters from analyses of large sets of agronomic data, but tools for visualizing data appear to be underutilized. Advances in color computer displays, color printing and software to create and manipulate images offer researchers numerous options for viewing data (Tufte, 1983; Fayyad et al., 2002; Keim, 2002). Data visualization with large data sets can facilitate identification of unexpected relations, potentially faulty data or other patterns. Visualization also provides a means for presenting complex results to nonspecialists in a manner that can facilitate comprehension.

Maps and data tables are familiar forms of data representation in agricultural and environmental science. Where applicable, maps are usually more effective for presenting data than tabular formats (Smelcer and Carmel, 1997). Geographic maps require a coordinate system related to geographic location, but for data with no inherent geographic structure, "pseudomaps" may be produced by using an artificial coordinate system. Data arrays are multi-dimensioned sets of data. Three-dimensional arrays have an obvious analogy with maps, where each *x*, *y*, *z* triplet in an array corresponds to the *x*- and *y*-coordinates of a point with an associated value *z*. In field research, many data sets have an array structure, where the genotype and environment are the *x*- and *y*-elements and the phenotypic value is the *z*-element. In molecular biology, the array concept is applied in laboratory assays using "microarrays," where binding agents (e.g., antibodies or oligonucleotides) are arranged in a rectangular array on a solid support (Schena et al., 1998; Ekins and Chu, 1999).

This paper illustrates how data arrays presented as pseudo-maps can enhance presentation and analysis of genotypic, phenotypic and environmental data in a single figure. These arrays are termed "GenPhEn arrays" to indicate the three components being presented, i.e., genotype, phenotype and environment. In data visualization literature, terms for similar presentations of arrays include "table visualization" (Hoffman and Grinstein, 2002), "pixel-oriented techniques" (Keim and Kriegel, 1996), "dense pixel displays" (Keim, 2002) and "matrix visualization" (Sharan et al., 2003).

2. Materials and methods

The basic approach for generating a GenPhEn array is to code a set of genotypic, phenotypic and environmental data as an *x*, *y*, *z*-array, where *x* might indicate the genotype, *y* the environment and *z* the phenotype. The coordinate value used to represent genotypes can be based on cultivar type, yield rank or other criteria. Environments are located on another axis using site mean yield, elevation, latitude or similar criteria. The phenotypic value, *z*, is plotted as a function of *x* and *y*. For a single trait, the phenotypic value may be used, but to view multiple traits in a single figure, values are standardized over a suitable range, typically from 0 to 1. The value for each point is indicated in the figure by shading the plotted symbol using a color scheme keyed to the range of values (e.g., red to green to blue for a 0–1 interval as used in Fig. 1).

2.1. Data sources

Applications of GenPhEn arrays are illustrated with data from a physiological study on lines of common bean (*Phaseolus vulgaris* L.) that characterized variation in leaf thickness (White and Montes-R, 2005), a multi-location wheat trial conducted by the International Center for Maize and Wheat Improvement (CIMMYT; Fox et al., 1997) and simulations of global warming scenarios for common bean.

2.1.1. Leaf thickness in common bean

The leaf thickness data are from a study conducted at two field stations of the International Center for Tropical Agriculture (CIAT) in Colombia using 25 bean lines that represented the two major gene pools (Andean and Mesoamerican) and also varied for growth habit and grain size (White and Montes-R, 2005). At Palmira (3°30′N latitude, elevation 965 m), trials were conducted in 1992 and 1993, and at Popayan in 1992. All trials were arranged in triple lattice designs and were managed for near optimal growth. Four traits related to leaf thickness were measured at 20, 32 and 48 days after planting. They were leaf thickness (THK) (measured with a hand-held micrometer), leaf optical density (LOD) at 670 nm (measured with a hand-held "chlorophyllometer"; Design Electronics, Palmerston

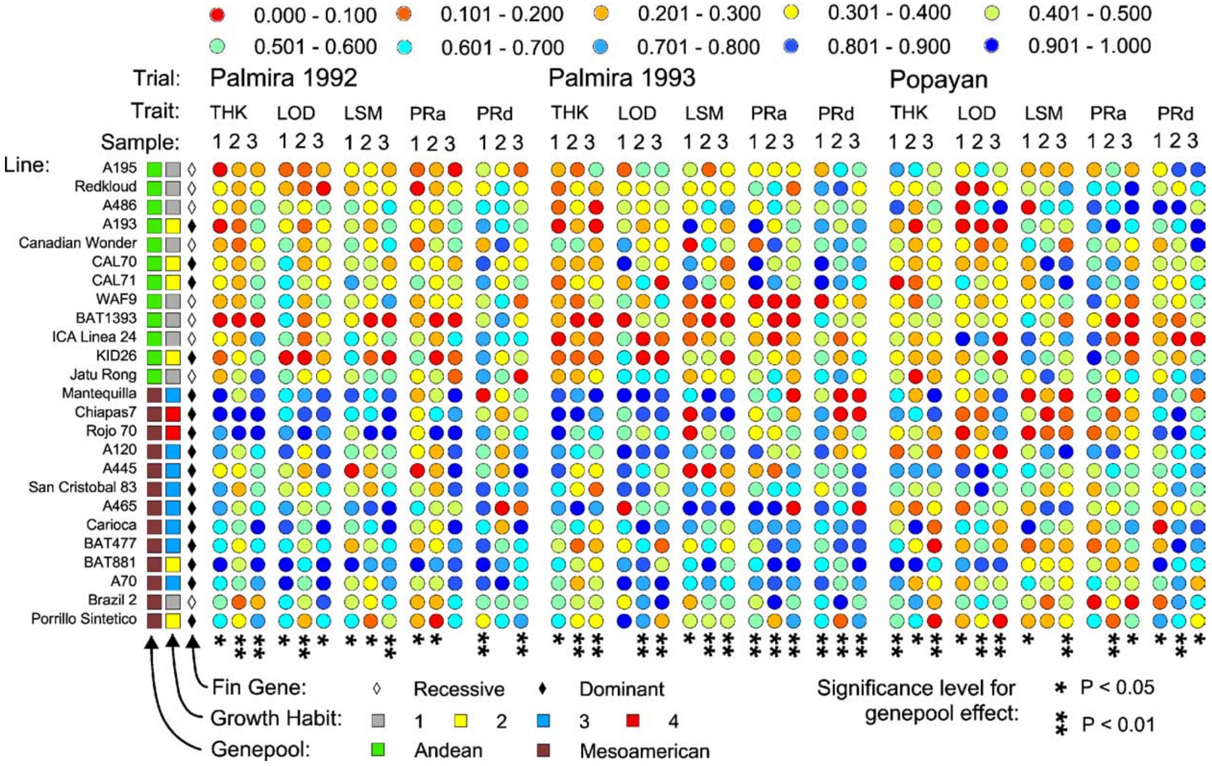

Fig. 1. GenPhEn array showing variation in five traits related to leaf thickness for 25 lines of common bean grown in three trials and sampled at 20, 32 and 48 days after planting. Values are presented in response classes based on standardized (0–1) values. Each line by sample by trial phenotypic value is the line mean. Values for each trait are displayed in the rectangularly blocked areas as labelled; where THK is thickness per second, LOD is leaf optical density, LSM is leaf specific mass, PRa is leaf protein content on an area basis and PRd is leaf protein content on a dry mass basis. Samples 1, 2 and 3 were taken 20, 32 and 48 days after planting, respectively. The symbols appended below the x-axis indicate significance levels of effects on genepool on each parameter x sample x trial combination. The three columns of points appended to the y-axis represent genepool, growth habit and dominant or recessive allelic nature for the *Fin* locus for each line. Growth habit was classified based on a 1–4 scale (Schoonhoven and Pastor-Corrales, 1987), where 1: determinate bush, 2: indeterminate bush, 3: sprawling indeterminate and 4: indeterminate climber (pole).

North, New Zealand[1]), leaf specific mass (LSM) (leaf dry mass per unit leaf area) and leaf protein content expressed on an area basis (PRa). Protein content on a dry weight basis (PRd) was also estimated. Details of crop management, measurements and data analyses are given in White and Montes-R (2005).

Data for means of individual lines for each trial and sample data was standardized to a range of 0–1, where 0 corresponded to the minimum value for each trait within a trial and sample date and 1 to the maximum value. The standardized phenotypic data were then grouped in an x, y, z array. The 25 genotypes were arranged by genepool and seed size within genepool on the y-axis. Data along the x-axis were grouped by trial, trait and sample date. The phenotypic value for each x, y-coordinate was colorized using a 10-color scale from red (minimum, 0) to blue (maximum, 1). The names of the 25 lines and classifications of lines by genepool, growth habit and the *Fin* genetic locus, which controls determinate and determinate stem type, were also displayed as columns of symbols along the y-axis.

2.1.2. *Elite spring wheat yield trial*

The 11th Elite Spring Wheat Yield Trial (ESWYT) included 30 lines (Table 1) and was grown at over 70

[1] Mention of a commercial or proprietary product does not constitute a recommendation by the USDA or the University of Georgia.

Table 1
Identification of lines tested in the 11th Elite Spring Wheat Yield Trial (ESWYT). Yield rank corresponds to position along the *x*-axis in Fig. 2

Name or cross	Mean yield ($\mathrm{kg\,ha^{-1}}$)	Yield rank	Days to heading (days)	Test weight ($\mathrm{g\,l^{-1}}$)	Kernel weight ($\mathrm{mg\,kernel^{-1}}$)
K134(60)/4/TOB/BMAN//BB/3/CAL	3710	1	80	762	40.0
Henne	3760	2	86	739	32.2
TR771773/SLM	3780	3	77	733	43.2
MAYA/NAC	3780	4	83	729	33.7
Rabe	3820	5	83	761	37.8
Caracara	3850	6	88	774	30.7
CNO79/PRL	3860	7	84	745	37.0
Fasan	3900	8	85	789	33.3
CNO79/PRL	3920	9	82	746	39.0
ARA	3960	10	80	756	36.1
Star	3960	11	90	749	39.7
Rayon F 89	3980	12	87	755	31.4
MRL/BUC	3400	13	83	746	37.3
Siren	4030	14	84	781	32.7
Fasan	4030	15	84	792	32.4
KA/NAC	4080	16	84	773	41.2
JUP/BJY	4090	17	83	781	32.1
Kauz	4130	18	87	774	31.8
Tepoca T 89	4140	19	84	764	37.0
Sibia	4150	20	88	750	41.1
Bacanora T 88	4150	21	86	771	32.5
Turaco	4170	22	83	758	40.5
PRL/VEE#6	4210	23	85	777	35.1
BJY/JUP	4220	24	87	766	32.6
Munia	4270	25	87	794	34.5
Sasia	4280	26	86	735	35.4
Angostura F 88	4300	27	87	771	39.6
Tui	4350	28	85	772	37.7
Veery	4370	29	86	737	35.4
Local check	4420	30	84	780	38.8
Minimum value from all trials	320		41	516	6.6
Maximum value from all trials	10870		182	867	59.0

sites (Table 2). Data were obtained from the International Wheat Information System, Version 2 (Fox et al., 1997). Trethowan et al. (2002) described general aspects of the objectives, experimental design and management of the ESWYT. The core data used for the GenPhEn illustration were genotypic means over replicates from individual environments (i.e., sites). Arrays were created for the traits of grain yield, days to heading, test weight (grain weight per unit volume) and kernel weight, using the 55 sites for which data for at least two traits of interest were available.

In order to combine arrays for the four traits, data were standardized to a range of 0–1, where 0 corresponded to the minimum value for each trait and 1 to the maximum value. Minimum and maximum values were calculated both across all sites and lines and for each site in order to present two different views of variability. The standardized phenotypic data for each trait were then grouped in an *x*, *y*, *z* array. The 30 genotypes were arranged from lowest to highest yield on the *x*-axis, the 55 sites from lowest to highest yield along the *y*-axis and the phenotypic value for each *x*, *y* coordinate colorized using a 10-color scale from red (minimum, 0) to blue (maximum, 1).

To illustrate how gene classification data can be appended to the *x*- and *y*-axes of a given array, the 30 lines were examined in terms of the alleles they possessed at each of four genetic loci, using data obtained from

Table 2
Locations represented in the 11th Elite Spring Wheat Yield Trial (ESWYT). Yield rank (from lowest to highest) corresponds to position along the *y*-axis in Fig. 2

Location	Country	Latitude (°)	Elevation (m)	Mean grain yield (kg ha^{-1})	Yield rank
Uyole Agriculture Center	Tanzania	−8.37	1801	1200	1
Taichung District Agriculture Improvement Station	Taiwan	26.53	15	1370	2
PBS Alentejo	Portugal	38.15	208	1470	3
Vollebekk	Norway	59.78	90	1680	4
Obonuco	Colombia	1.27	2710	1800	5
OCEPAR-Palotina	Brazil	−24.83	341	1880	6
Bordenave	Argentina	−37.02	212	2120	7
CNPT-EMBRAPA	Brazil	−28.42	684	2270	8
San Benito	Bolivia	−17.10	2730	2290	9
Heihe Agricultural Research Institute	China	50.45	168	2390	10
Jiangsu Academy of Agriculture Science	China	32.78	67	2500	11
FUNDACEP	Brazil	−28.60	473	2500	12
OR Melhoramento de Sementes Ltda.	Brazil	−28.43	700	2550	13
Labor Ovalle	Guatemala	14.50	2407	2610	14
Bembeke	Malawi	−14.43	1560	2650	15
Southeastern Anatolian Agricultural Research Institute	Turkey	37.20	660	2680	16
ICGR, Beijing	China	39.47	50	2866	17
Joydepur	Bangladesh	23.38	8	2879	18
Jiu San Agriculture Institute	China	48.28	288	2888	19
Small Grain Institute	South Africa	−28.20	1687	2985	20
NPBRC-Njoro	Kenya	0.00	2165	3000	21
Marcos Juarez	Argentina	−32.12	110	3050	22
CRIA	Paraguay	−27.82	200	3070	23
Centro Experimental Chimaltenango	Guatemala	14.83	1790	3160	24
Tomejil	Spain	37.58	72	3210	25
Tibiatata	Colombia	4.20	2550	3360	26
Sokolac Centre	Yugoslavia	43.80	860	3380	27
Arusha Farm, T.A.R.O.	Tanzania	−3.85	1372	3400	28
Pirsabak	Pakistan	33.98	340	3480	29
Hissar	India	29.77	215	3520	30
Institute of Hongxinglong	China	46.55	75	3680	31
Andenes-Cusco Anexo Taray	Peru	−13.87	2900	3740	32
Crop Science Institute Sichuan Academy of Agriculture Science	China	30.02	506	4080	33
Durgapura	India	26.80	450	4110	34
NIAB Faisalabad	Pakistan	31.08	117	4200	35
PAU-Ludhiana	India	30.87	247	4210	36
The Keshan Institute of Heilong Jiang Academy	China	48.88	223	4410	37
Ahwaz	Iran	31.67	21	4530	38
L'Urgell (Palau D'Anglesola)	Spain	41.92	250	4540	39
Kentziko Thermi	Greece	40.95	10	4640	40
Azad University of Agriculture Techology Kanpur	India	26.40	123	4780	41
Araghi Mahaleh	Iran	36.33	5	5110	42
Rncho del la Merced (Jerez)	Spain	36.15	20	5290	43
EMBRAPA-CPAC	Brazil	−15.70	1001	5310	44
Rattray Arnold Research Station	Zimbabwe	−17.23	1300	5450	45
Fars	Iran	28.55	1101	5810	46
Jesus Ma. Jalisco	Mexico	20.18	2110	5820	47
Graneros	Chile	−34.70	479	5960	48

Table 2 (*Continued*)

Location	Country	Latitude (°)	Elevation (m)	Mean grain yield (kg ha^{-1})	Yield rank
Mimosa	Madagascar	−20.00	1501	6790	49
Spii Cereal Research Station	Iran	35.97	1321	7670	50
Quilamapu	Chile	−36.92	217	8480	51
Wang Tai Pu	China	38.25	1118	8570	52
Dirab	Saudi Arabia	24.62	600	9000	53
La Platina	Chile	−33.63	629	9270	54
Kufra Production Project	Libya	25.00	415	9360	55

the on-line database Wheat Pedigree and Identified Alleles of Genes (Martynov et al., 2002). The four loci were the leaf rust resistance genes *Lr13* and *Lr26* and the height reducing genes *Rht1* and *Rht2*. Lines known to be homozygous for the dominant or recessive allele were coded 1 (blue) or 0 (red), respectively, for the given locus. If there was no information for allele status of a line, the *x*- and *y*-position was left blank. The gene classification data for each line were thus appended as a 2^4 factorial array.

The sites used in the 11th ESWYT were also characterized by computing standardized values for site elevation (in this case from 5 to 2900 m) and site latitude (absolute values from 0° to 60° latitude), again scaled and coded using the 0–1 scale. The values were appended as two columns of points along the *y*-axis.

2.1.3. Simulations of effects of temperature regime on common bean

The simulation study examined the potential impact of global warming on rainfed bean production in Michigan as simulated by the CSM-GeneGro model (Hoogenboom et al., 2004). CSM-GeneGro is a process-based model that simulates photosynthesis, respiration, partitioning and dynamics of water and nutrients. It combines features from the previously developed GeneGro model (White and Hoogenboom, 1996; Hoogenboom et al., 1997; Hoogenboom and White, 2003) with the Cropping System Model (CSM) (Jones et al., 2003). For the GenPhEn array, a hypothetical congenic set of cultivars was created that represented all possible homozygous combinations of the alleles available at each of seven genetic loci (Table 3). Genotypes for each locus were coded with a value of 0 for recessive and 1 for dominant, and then linear equations were used to estimate the actual coefficients used in the simulations. Of 128 possible different homozygous genotypic combinations (2^7), only 96 different phenotypic combinations are possible since in a *ppd* homozygote, the *Hr* and *hr* homozygotes have undistinguishable phenotypes (Kornegay et al., 1993). Details of the procedure used to develop the estimator equations based on field observations of breeding lines and cultivars are given in White and Hoogenboom (1996). Daily weather data from 1930 to 2002 were obtained for the Kellogg Biological Station (42.40° North, 85.38° West; elevation 277 m). The soil was a Kalamazoo Loam (Typic Hapludalf), allowing root growth to a depth of 0.9 m. Crops were planted on 18 June at a population of 25 plants m^{-2} and a 0.6 m row width. The crop was rainfed and 50 kg ha^{-1} of nitrogen was applied at planting.

Possible effects of global warming were simulated by increasing all daily maximum and minimum temperatures in 1 °C increments up to a 4 °C increase above the historic values. No modifications were made to radiation or precipitation regimes or to management.

For each of the 96-genotypes, means for various crop traits and environmental factors were calculated over the 73-year-period. Duration of grain fill was estimated as the difference between anthesis and matu-

Table 3
Effects of genes in common bean that are considered in the CSM-GeneGro crop simulation model

Gene	Effect of dominant allele
Ppd	Long days delay flowering (classic short-day response)
Hr	Increases effect of *Ppd*, but requires *Ppd* to be present
Fin	Indeterminate stem type, which is associated with later flowering
Fd	Early flowering and maturity
Ssz1	Increases seed size
Ssz2	Increases seed size
Ssz3	Increases seed size

Table 4
Mean minimum and maximum values obtained for the simulation of 96-genotype combinations of rainfed common bean crops over 73 years and five temperature regimes at the Kellogg Biological Station, Michigan

Variable	Units	Minimum value	Maximum values
Grain yield	kg ha^{-1}	680	1850
Canopy dry weight	kg ha^{-1}	1500	4230
Days to anthesis	days	29	49
Duration of grain filling	days	30	42
Grain weight	mg	100	360
Harvest index	kg ha^{-1}	32	55
Total season precipitation	mm	560	660
Total season evapotranspiration	mm	440	530
Water use efficiency	kg ha^{-1} mm^{-1}	1.4	3.6
Season nitrogen uptake	kg ha^{-1}	43	66
Nitrogen use efficiency	kg ha^{-1}	12.4	30.4

rity dates. Water use efficiency (WUE) and nitrogen use efficiency (NUE) were calculated as the ratio of crop grain yield to total evapotranspiration and crop nitrogen uptake, respectively. For plotting, the crop trait values and environmental factor values were standardized on a 0–1 scale as applied to the respective genotypes and temperature regimes. In contrast to the wheat data set, these data were standardized on a 0–1 scale as applied to the respective minimum and maximum observed over all temperature regimes (Table 4). To illustrate an alternative means of positioning the 96-genotypes along the *x*-axis, phenotypic values for days to anthesis, duration of grain-filling, canopy weight at maturity, grain yield, kernel weight and harvest index from simulations under the historic temperature regime were clustered using Ward's minimum variance procedure (SAS Institute, 1996). The resulting six clusters were then used to order the display of genotypes along the *x*-axis of the array, with the within cluster order being based on the relative order used in the previous array.

2.2. Data processing and plotting

Initial data processing to calculate standardized values, obtain means and ranks, and generate *x*- and *y*-coordinates was done with the SAS System for Windows, Release 8.00 (SAS Institute Inc., Cary, IN). Draft GenPhEn arrays were produced with the GPLOT procedure of SAS, and final published forms with more detailed labeling were composed using ESRI ArcMap 8.2 (Environmental Systems Research Institute Inc., Redlands, CA).

3. Results and discussion

3.1. Leaf thickness in common bean

Bean cultivars differ in leaf thickness and this variation is associated with the genepool of origin. Materials from the Mesoamerican genepool have thicker leaves, which is associated with greater leaf assimilation rates, relative growth rate and seed yield (Sexton et al., 1997; White and Montes-R, 2005). The traits THK and LOD show promise for screening genotypes for leaf thickness, but large interactions of lines with trials and sample dates can occur (White and Montes-R, 2005). Andean lines are needed that consistently show high values of thickness parameters across samples, seasons and locations, and thus might serve as parents to develop Andean germplasm with thicker leaves.

In Fig. 1, the trend of Andean lines having thinner leaves is apparent for THK, LOD, LSM and PRa for both years at Palmira but less so at Popayan. PRd also varied with genepool. Andean lines that had higher values for the thickness traits included A 486, CAL 71and Jatu Rong, while BAT 1393 had extremely thin leaves. This ranking could also be determined by examining means over standardized values of all traits, but the large interactions for line with trial and sample date imply that overall means may be misleading. In the array, interactions are indicated by irregular patterns of low versus high values. One example is for THK in Palmira 1992, where several Andean materials had values similar to the Mesoamerican lines in sample 3 (48 DAP) but not sample 1 (20 DAP) and 2 (32 DAP).

Another is the large number of high values of PRa for Andean lines in Popayan as compared to the two trials at Palmira. Thus, revised criteria for promising lines might exclude PRa and any data from 48 DAP. Using these criteria, CAL 71, CAL 70 and Jatu Rong have the thickest leaves based on the standardized values. The information on growth habit and the *Fin* gene shows that these lines include indeterminate and determinate types, suggesting that leaf thickness is not confounded with growth habit.

The array for leaf thickness traits (Fig. 1) thus serves as a valuable platform for interactive data exploration and hypothesis generation. For simplicity, we only calculated means using different criteria. In more extensive analyses, statistical tests could assess whether partitioning data into subsets of lines, locations or samples reduced the interactions and clarified underlying patterns.

3.2. Elite spring wheat yield trial

Two GenPhEn arrays from the 11th ESWYT are presented in Fig. 2. In both plots, the *x*-axis corresponds to the mean yield (ranked from lowest to highest) of the 30 lines averaged over the relevant number of test sites, with the four horizontally arrayed data blocks corresponding to the four measured traits. Similarly, the *y*-axis corresponds to the mean yield (ranked from lowest to highest) of the given test site, averaged over the 30 lines. In Fig. 2A, the standardized values are based on the minimum–maximum range of values using line means averaged over all locations. The trend of the lowest grain yields appearing at the bottom (red color) and the highest yields at the top (blue) reflects the sorting of *y*-values using site mean yield. Deviations from this trend imply a genotype *x*-site interaction. For example, site 21 on the *y*-axis (NPBRC-Njoro, Kenya) stands out for having two high-yielding lines amongst the otherwise medium to low-yielding lines. Note that line 1, whose position along the *x*-axis identifies it as having the lowest mean yield averaged over all sites, had a reported yield of more than 10,000 kg ha^{-1} at test site 21. In contrast, line 30 had the highest mean yield averaged over sites, but at site 28 (Arusha Farm, T.A.R.O., Tanzania), it yielded 1800 kg ha^{-1} as compared to yields of over 4000 kg ha^{-1} obtained for six lines. Line 30 was the “local check,” which was not the same genotype at all test sites but usually is a local cultivar that provides a benchmark to determining whether new lines have suitable yields or other traits at the test site in question.

For days to heading, test weight and kernel weight (Fig. 2A), the lack of a similar pattern of color banding from low- to high-grain yield suggests no simple relationship between these traits and grain yield. Nonetheless, for days to heading, if the two sites with the largest number of days to heading (sites 11 and 16) are ignored, the upper half of the days to heading data block has fewer red points, suggesting that sites allowing longer vegetative growth had higher yields.

For all traits in Fig. 2A, little variation in color within a row indicates little difference in the performance of the lines at the corresponding site. Conversely, marked differences in color within a row indicate that lines varied in performance. Thus, for genotypic variation in test weight, sites 6 (OCEPAR-Palotina) and 14 (Labor Ovalle) might be of special interest, whereas for kernel weight, sites 4 (Vollebekk), 13 (Or. Melh. de Sementes) and 34 (Durgapura) produced large ranges of variation. Such patterns can also be discerned in tables of data and accompanying statistical tests, but for large numbers of variables and treatments, a GenPhEn array greatly facilitates identifying interesting patterns as well as potential problem data.

Standardizing the phenotypic data based on minimum and maximum values at each test site emphasizes variation of lines within sites (Fig. 2B). Line 3 (Henne), while having a low-mean yield over test sites, was the highest yielding line at sites 48 and 53. In contrast, while line 29 (a selection of Veery) was the second highest yielding over test sites, it performed poorly at sites 17 (ICGR, Beijing) and 24 (Centro Exp. Chimaltenango). Line 25 (Munia) stood out for relatively high yields at lower yielding sites. For days to heading, line 3 was consistently the earliest and line 11 (Star) was consistently the latest. Line 3 also had low-test weights but high-kernel weights. Often, the local check (line 30) combined moderately high-test weight with high-kernel weight.

The elevation and latitude of the test sites, as indicated by the columns of points, did not show a strong relation with the four traits (Fig. 2A and B), but the local information is of use for characterizing individual sites. Thus, site 4 (Vollebekk) is identifiable as a high-latitude, low-elevation site and site 5 (Obonuco) as a low-latitude, high-elevation site.

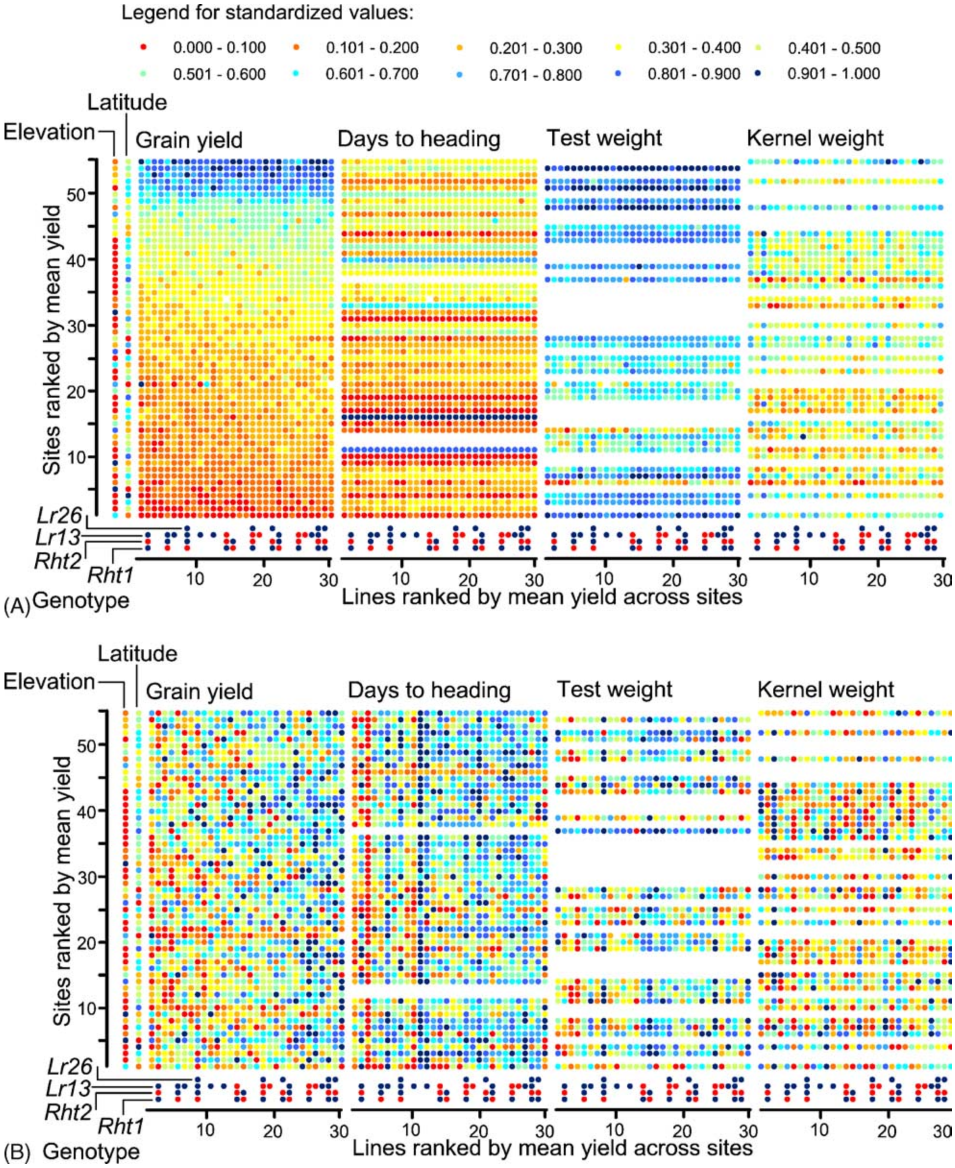

Fig. 2. GenPhEn arrays of response classes based on standardized (0–1) grain yield, days to heading, test weight and kernel weight from the 11th ESWYT (units and minimum–maximum ranges are given in Table 1). Each test site by line phenotypic value is by line mean yield averaged over all sites on the *x*-axis and by site mean yield averaged over lines on the *y*-axis. Values for each trait are displayed in the rectangularly blocked areas as labeled. Appended below the *x*-axis is a four-row factorial array denoting homozygous dominant or recessive allelic nature of the given lines for the *Lr26*, *Lr13*, *Rht2* and *Rht1* loci (blank where data were unavailable). The two columns of points appended to the *y*-axis represent standardized values for the elevation and latitude of each test site. If a trait was not measured at a given test site, the data row is blank. (A) Standardized values based on minimum–maximum ranges of phenotypic values averaged over all sites. (B) Standardized values based on minimum–maximum ranges of phenotypic values within each test site.

The arrays of ESWYT data (Fig. 2A and B) showed that the approach easily scales to larger amounts of data, including when substantial blocks of data are missing. The arrays again served to identify extreme values (data exploration) and to suggest patterns (hypothesis generation) that could be examined further with regression analyses or other methods.

3.3. Simulations of effects of temperature regime on common bean

In modeling response of common bean to increased air temperature, 11 variables were plotted as parallel rectangles of points, with the 96-genotype combinations indicated by the rectangle below the main data (Fig. 3). Within each rectangle for a variable, the five temperature regimes correspond to the five rows of points. Simulated grain yield, canopy dry weight and grain weight predominantly declined with increasing temperature (Fig. 3A).

For simulations using historic weather conditions (the lowermost row in each trait rectangle of Fig. 3A and B), the differential effects of the seven genes are most apparent for grain weight and days to anthesis. Clustering improves the visualization with respect to genes affecting those traits (Fig. 3B). Note that early-flowering genotypes are predominantly *Fd Fd*, whereas the groups of small-seeded genotypes are *ssz1 ssz1*. The highest yielding genotypes are mainly *Ppd Ppd*.

With warming to +4 °C, yields decrease more than 50% (Fig. 3). The duration of vegetative growth, indicated by time to anthesis, is reduced more than grain filling duration. Shortening of the growth cycle is reflected in low-canopy dry weights, so part of the yield decline is attributable to reduced growth. Less efficient partitioning to yield at higher temperatures is also implicated, however by the decline in harvest index.

Season totals for precipitation and evapotranspiration showed little variation with temperature regime, and thus WUE largely varied with grain yield. Nitrogen uptake decreased with increasing temperature, but this again appeared linked to overall growth. However, due to the season length effect noted for grain yield, NUE for some genotypes was highest with a warming of 1–2 °C.

The clusters (Fig. 3B) showed unexpected heterogeneity for genotypes. Rather than grouping results by genotypes, clusters 1, 2 and 3 contained similar proportions of dominant and recessive genotypes. Cluster 4 consisted of predominantly *Ppd Ppd Fin Fin fd fd* genotypes, which had later flowering and lower harvest index (Fig. 3B). In contrast, clusters 5 and 6 were exclusively *fin fin Fd Fd*, making them determinate and early to flower and mature. A tentative conclusion is that different combinations of alleles can result in phenotypes that are similar for the traits considered.

3.4. Comparisons among the three examples

For three different types of data, the GenPhEn arrays provided a comprehensible means for summarizing and presenting large amounts of data (400 points in Fig. 1 and approximately 5300 points in each plot of Figs. 2 and 3) in a compact yet meaningful manner. Varying how data were standardized (Fig. 2A versus B) or grouped (Fig. 3A versus B) allowed the data to be examined from different perspectives. Numerous other options exist for enhancing the information in the arrays. Using higher resolution graphics, one can display data points with different symbols to indicate levels of significance or additional traits of interest, a technique termed “iconic display” by Keim (2002). Various transformations or alternate strategies for normalization could also be employed. More advanced clustering routines such as biclustering (two-way clustering) might assist pattern detection in larger arrays.

Using space adjacent to the main array to display additional information on lines, sites or levels of significance can further enhance interpretations. In the study on leaf thickness, the effect of genepool on the traits was readily seen by comparison to the genepool classification and further confirmed by the significance tests displayed below the array. For the bean simulations, indexing by the seven-locus factorial set of genotypes facilitated interpretation. Indexing could also be used for other traits, including key disease resistance or quality traits that show low genotype *x* environment effects. Similarly, in Fig. 2, characteristics of the ESWYT sites (i.e., latitude and elevation) were plotted to the left of the main array. This approach could be extended to climatic or edaphic conditions (e.g., mean temperature or initial soil mineral nitrogen) or trial management (use of irrigation or pesticides) could be displayed to the left of the *y*-axis. Alternatively, principal components could be estimated for combinations of genotype, environment or genotype by environment effects and used

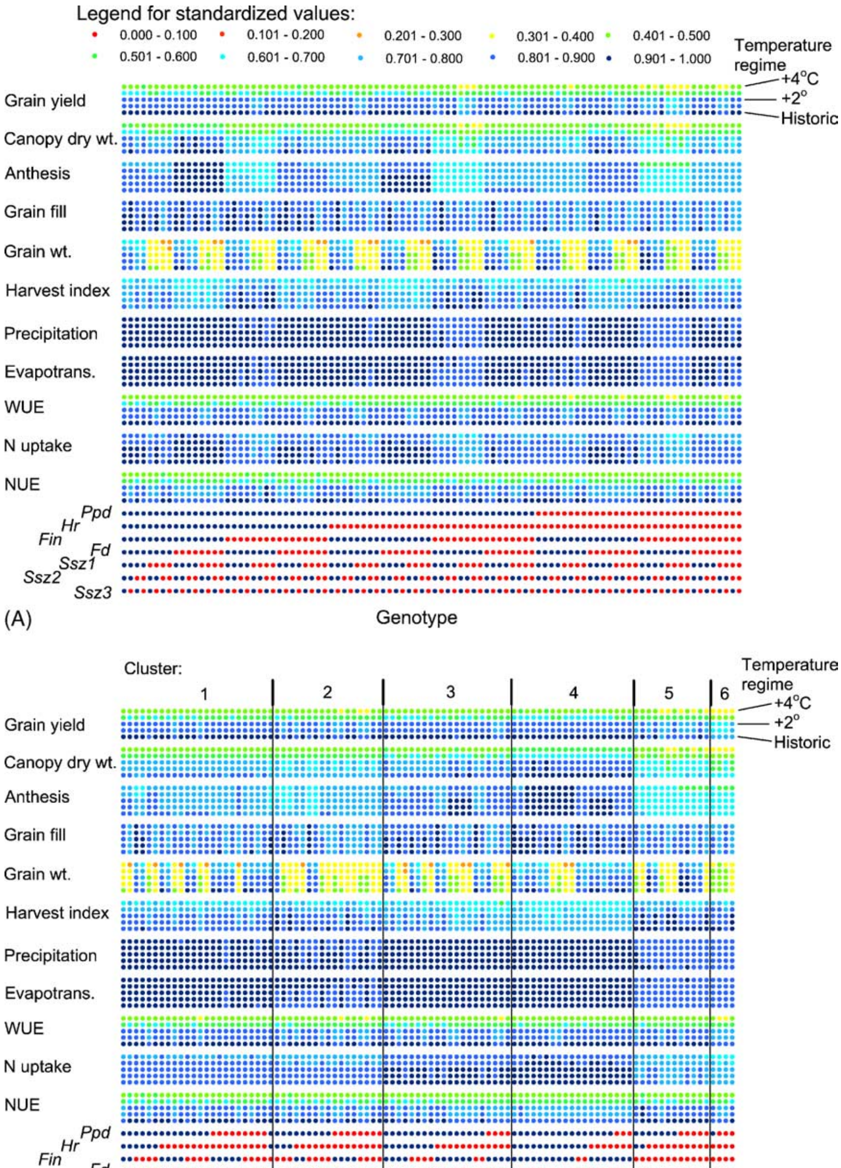

Fig. 3. GenPhEn arrays of response classes based on standardized (0–1) for grain yield, canopy dry weight at maturity, days to anthesis, duration of grainfilling, grain weight, harvest index, season precipitation, season evapotranspiration, water use efficiency, season nitrogen uptake and nitrogen use efficiency as simulated with CSM-GeneGro. Values are given for 96 distinct phenotypes possible from 128 homozygous combinations of seven genes (2^7) in common bean (units and minimum–maximum ranges are given in Table 4), and field conditions are for rainfed production at the Kellogg Biological Station, Michigan. Each row within a trait grouping represents one of five temperature regimes, ranging from the average historic value (1930–2002) to a +4 °C increase. Appended below the *x*-axis is a seven-row factorial array denoting homozygous dominant or recessive allelic nature of the 96-genotypes for the *Ppd*, *Hr*, *Fin*, *Fd*, *Ssz1*, *Ssz2* and *Ssz3* loci, with reach row corresponding to a different locus and values of 0 or 1, to recessive or dominant alleles. (A) Genotypes ordered by gene combinations. (B) Genotypes classified into six clusters as indicated along the top of the array.

to order the data along both the *x*- and *y*-axes, allowing the GenPhEn arrays to be used in conjunction with biplots (e.g., Yan et al., 2001).

4. Conclusion

GenPhEn arrays seem especially useful for exploratory data analysis and hypothesis generation, and their use is recommended as a complement to quantitative statistical analyses. We emphasized examining phenotypes, treating genotypes and environments as secondary data, but the approach of plotting arrays as pseudo-maps is directly applicable to research where the focus is on characterizing gene expression or environmental variation. For gene expression, the approach converges with displays of data from microarrays and other tools of bioinformatics (e.g., Eisen et al., 1998; Getz et al., 2000; Sharan et al., 2003).

The GenPhEn arrays were produced with the GPLOT procedure of SAS/GRAPH and with ArcMap, but other software packages have similar capabilities. GPLOT offered the convenience of producing output within the same programming environment as the initial data analysis. ArcMap offered greater power for composing figures and for interactive querying, but it required that the raw input include pre-calculated *x*- and *y*-coordinates. For simplicity, the examples presented did not take full advantage of high-resolution graphics nor of interactive querying. Numerous modifications in terms of groupings of data use of special symbols and indexing can also be envisaged. Sample SAS programs for generating and displaying GenPhEn arrays are available from the senior author.

Acknowledgments

The authors thank Bruce Kimball for comments on map composition and James Specht, Graham McLaren and Paul Pinter for suggestions on manuscript content.

References

Collaku, A., Harrison, S.A., Finney, P.L., Van Sanford, D.A., 2002. Clustering of environments of southern soft red winter wheat region for milling and baking quality attributes. Crop Sci. 42, 58–63.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. PNAS 95, 14863–14868.

Ekins, R., Chu, F.W., 1999. Microarrays: their origins and applications. Trends Biotech. 17, 217–218.

Fayyad, U., Grinstein, G.G., Wierse, A. (Eds.), 2002. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, San Francisco, 407 pp.

Fox, P.N., Magaña, R.I., Lopez, C., Sanchez, H., Herrera, R., Vicarte, V., White, J.W., Skovmand, B., Mackay, M.C., 1997. International Wheat Information System (IWIS), Version 2 [CD-ROM publication]. CIMMYT, Mexico, D.F.

Getz, G., Levine, E., Domany, E., 2000. Coupled two-way clustering analysis of gene microarray data. PNAS 97, 12079–12084.

Hoffman, P.E., Grinstein, G.G., 2002. A survey of visualizations for high-dimensional data mining. In: Fayyad, U., Grinstein, G.G., Wierse, A. (Eds.), Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, San Francisco, pp. 47–82.

Hoogenboom, G., White, J.W., 2003. Modification of a crop simulation model that incorporates gene action. Agron. J. 95, 82–89.

Hoogenboom, G., White, J.W., Messina, C., 2004. From genome to crop: integration through simulation modeling. Field Crops Res. 90, 145–163.

Hoogenboom, G., White, J.W., Acosta-Gallegos, J., Gaudiel, R., Myers, J.R., Silbernagel, M.J., 1997. Evaluation of a crop simulation model that incorporates gene action. Agron. J. 89, 613–620.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

Keim, D.A., Kriegel, H.-P., 1996. Visualization techniques for mining large databases: a comparison. IEEE Trans. Knowl. Data Eng. 8 (6), 1–29.

Keim, D.A., 2002. Information visualization and visual data mining. IEEE Trans. Vis. Comp. Graph 8 (1), 1–8.

Kornegay, J., White, J.W., Dominguez, J.R., Tejada, G., Cajiao, C., 1993. Inheritance of photoperiod response in Andean and Mesoamerican common bean. Crop Sci. 33, 977–984.

Lin, C.S., Binns, M.R., Lefkovitch, L.P., 1986. Stability analysis: where do we stand? Crop Sci. 26, 894–900.

Martynov, S.P., Dobrotvorskaya, T.V., Hon, I., Faberova, I., 2002. Wheat Pedigree and Identified Alleles of Genes (Online). Available at http://genbank.vurv.cz/wheat/pedigree/default.htm (posted 8 April 2002; verified 12 October 2004).

SAS Institute, 1996. SAS User's Guide Statistics. SAS Institute, Cary, NC.

Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E., Davis, R.W., 1998. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotech. 16, 301–306.

Schoonhoven, A.V, Pastor-Corrales, M.A., 1987. Standard System for the Evaluation of Bean Germplasm. CIAT, Cali, Colombia, 56.

Sexton, P.J., Peterson, C.M., Boote, K.J., White, J.W., 1997. Early-season growth in relation to region of domestication, seed-size, and leaf traits in common bean. Field Crops Res. 54, 163–172.

Sharan, R., Maron-Katz, A., Shamir, R., 2003. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Bioinformatics 19, 1787–1799.

Smelcer, J.B., Carmel, E., 1997. The effectiveness of different representations for managerial problem solving: comparing tables and maps. Decis. Sci. J. 28, 391–420.

Trethowan, R.M., van Ginkel, M., Rajaram, S., 2002. Progress in breeding wheat for yield and adaptation in global drought affected environments. Crop Sci. 42, 1441–1446.

Tufte, E.R., 1983. The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut, 197 pp.

White, J.W., Hoogenboom, G., 1996. Simulating effects of genes for physiological traits in a process-oriented crop model. Agron. J. 88, 416–422.

White, J.W., Montes-R, C., 2005. Variation in parameters related to leaf thickness in common bean (*Phaseolus vulgaris* L.). Field Crops Res. 91 (1), 7–21.

White, J.W., Singh, S.P., Pino, C., Rios B., M.J., Buddenhagen, I., 1992. Effects of seed size and photoperiod response on crop growth and yield of common bean. Field Crops Res. 28, 295–307.

Yan, W., Cornelius, P.L., Crossa, J., Hunt, L.A., 2001. Two types of GGE biplots for analyzing multi-environment trial data. Crop Sci. 41, 656–663.